**NEW EVOLUTIONARY LEARNING APPROACHES FOR FEATURE SELECTION ON BIOMEDICAL DATA**

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

**BACHELOR OF TECHNOLOGY**

IN

**COMPUTER SCIENCE ENGINEERING**

SUBMITTED BY

|  |  |
| --- | --- |
| **Name** | **Univ. Roll No.** |
| Aparna Mahato | 10800118116 |
| Anwesha Dey | 10800118117 |
| Sanraj | 10800118056 |
| Rahul Prasad | 10800118074 |
|  |  |

UNDER THE GUIDANCE OF

**MR. SABYASACHI MUKHERJEE**

Associate Professor, Department of CSE



DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

**ASANSOL ENGINEERING COLLAGE**

AFFILIATED TO

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY

January, 2022

|  |  |
| --- | --- |
|  | **DEPARTMENT OF COMPUTER SCIENCE ENGINEERING** |
| **ASANSOL ENGINEERING COLLEGE**  Vivekananda Sarani, Kanyapur, Asansol, West Bengal – 713305 |

*Certificate of Recommendation*

I hereby recommend that the preliminary thesis report entitled, **“New Evolutionary learning approaches for feature selection on biomedical data.”** carried out under my supervision by the group of students listed belowmay be accepted in partial fulfilment of the requirement for the degree of “Bachelor of Technology in **COMPUTER SCIENCE** Engineering” of Asansol Engineering College under MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY.

|  |  |
| --- | --- |
| **Name** | **Univ. Roll No.** |
| Aparna Mahato | 10800118116 |
| Anwesha Dey | 10800118117 |
| Sanraj | 10800118056 |
| Rahul Prasad | 10800118074 |
|  |  |

…………………………………

(Mr. Sabyasachi Mukherjee)

Thesis Supervisor

A

Dept. of Computer Science Engineering,

Asansol Engineering College,

Asansol-713305

Countersigned:

………………………………

(Dr. Debasish Chakraborty)

Head of the Department

Dept. of Computer Science Engineering,

Asansol Engineering College,

Asansol-713305

|  |  |
| --- | --- |
|  | **DEPARTMENT OF COMPUTER SCIENCE ENGINEERING** |
| **ASANSOL ENGINEERING COLLEGE**  Vivekananda Sarani, Kanyapur, Asansol, West Bengal – 713305 |

***Certificate of Approval***

The forgoing thesis is hereby approved as creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it is submitted.

…………………………………

(**Mr. Sabyasachi Mukherjee**)

Thesis Supervisor

Dept. of Computer Science Engineering,

Asansol Engineering College,

Asansol-713305

***Acknowledgement***

It is our great privilege to express our profound and sincere gratitude to our Project Supervisor, **Mr. Sabyasachi Mukherjee** for providing us a very cooperative and precious guidance at every stage of the present project work being carried out under his/her supervision. His valuable advice and instructions in carrying out the present study has been a very rewarding and pleasurable experience that has greatly benefited us throughout the course of work.

We would like to convey our sincere gratitude towards **Dr. Debasish Chakraborty**, Head of the Department of **Computer Science Engineering**, Asansol Engineering College for providing us the requisite support for timely completion of our work. We would also like to pay our heartiest thanks and gratitude to all the teachers of the Department of **Computer Science Engineering**, Asansol Engineering College for various suggestions being provided in attaining success in our work.

We would like to express our earnest thanks to **Mr. Suman Mallick**, of CSE Project Lab for his technical assistance provided during our project work.

Finally, I would like to express my deep sense of gratitude to my parents for their constant motivation and support throughout my work.

…………………………………

Aparna Mahato

(Student 1)

…………………………………

Anwesha Dey

(Student 2)

…………………………………

Sanraj

(Student 3)

…………………………………

Rahul Prasad

(Student 4)

***Abstract***

In this era of technology and innovation, human needs are rapidly increasing and new ways to improve the efficiency or to ease the process are constantly being invented. Human needs have no bound, but our bodies do have constraints. In an effort to diminish these constraints these following studies are performed. High dimensional biomedical data contain tens of thousands of features, accurate and effective identification of the core features in these data can be used to assist diagnose related diseases. However, there are often a large number of irrelevant or redundant features in biomedical data, which seriously affect subsequent classification accuracy and machine learning efficiency.

To solve this problem, a novel filter feature selection algorithm based on redundant removal (FSBRR) of Machine Learning is proposed to classify high dimensional biomedical data in this paper. First of all, two redundant criteria are determined by vertical relevance (the relationship between feature and class attribute) and horizontal relevance (the relationship between feature and feature). Secondly, to quantify redundant criteria, an approximate redundancy feature framework based on mutual information (MI) is defined to remove redundant and irrelevant features. To evaluate the effectiveness of our proposed algorithm, controlled trials based on typical feature selection algorithm are conducted using three different classifiers, and the experimental results indicate that the FSBRR algorithm can effectively reduce the feature dimension and improve the classification accuracy.

In addition to that, different types of Model training algorithms are used to diagnose the large Dataset collected for Predicting Cancerous Cells in Breasts, which are then tested on an unknown dataset for calculating the accuracy.

Further, this trained model can be used an automated assistance for medical staffs to recognise and control the growth of cancerous cells in humans. Thus, bringing an evolution to the world of biomedical sciences.

***Contents***

|  |  |  |  |
| --- | --- | --- | --- |
| Certificate of Recommendation……………………………………………………. | | |  |
| Certificate of Approval……………………………………………………………... | | |  |
| Acknowledgement………………………………………………………….............. | | |  |
| Abstract………………………………………………………………………………  Lists of Figures……………………………………………………………………… | | |  |
| **Contents……………………………………………………** | | |  |
|  | | |  |
| **1.** | **Preface……………………………………………………………………** | |  |
|  | 1.1 | Introduction…………………………………………………….. | 1 |
|  | 1.2 | Motivation of the project………………………………….…… | 1 |
|  | 1.3 | Basic description of the project………………………….…………… | 1 |
| **2.** | **Literature Review………………………………………………………………** | |  |
|  |  | Review of related works …………………………. ………………… | 3 |
| **3.** | **Proposed model/algorithm…………………………….………………………** | |  |
|  |  | Proposed algorithms/ models………………………………………. | 5 |
| **4.** | **Simulation and Results…………………………………………………………** | |  |
|  | 4.1 | Experimental set up …………………………………..……………. | 8 |
|  | 4.2 | Experimental results………...……………...……………………….. | 9 |
| **5.** | **Discussion and Conclusion ……...…………………………………….……….** | |  |
|  | 5.1 | Discussion…………………………………..…………………….… | 15 |
|  | 5.2 | Future work…………………………………..………………...…….. | 15 |
|  | 5.3 | Conclusion…………………………………………………………….. | 15 |
|  | **References………………………………………………………………………..** | | 16 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| ***List of Figures*** | | |  |

Fig. 3.1 Logistic Regression Graph 5

Fig. 3.2 Decision Tree Process 6

Fig. 3.3 Feature Selection Uses 6

Fig. 3.4 Random Forest Classifier 7

Fig. 4.1 Workflow Process 9

Fig. 4.2.1 Module Import and Dataset Loading 9

Fig. 4.2.2 Feature Selection, Checking for Null Values 10

Fig. 4.2.3 Data Cleaning and Counting Classifying Variables 10

Fig. 4.2.4 Plotting Bar Graphs for Classifying Variables 10

Fig. 4.2.5 Data Preprocessing 11

Fig. 4.2.6 Pairplot is Plotted Against Each Pair of Features 11

Fig. 4.2.7 Correlation Matrix 12

Fig. 4.2.8 HeatMap on Correlation Matrix 12

Fig. 4.2.9 Splitting Dataset into Training and Test Dataset 13

Fig. 4.2.10 Model Training on Training Dataset 13

Fig. 4.2.11 Confusion Matrix and Accuracy 13

Fig. 4.2.12 Precision 14

1. PREFACE

1.1 Introduction:

Breast cancer is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society.

The early diagnosis of Breast Cancer can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors and malignant tumors can prevent patients undergoing unnecessary treatments.

Thus, the correct diagnosis of Breast Cancer and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex Breast Cancer datasets, machine learning (ML) is widely recognized as the methodology of choice in Breast Cancer pattern classification and forecast modelling.

Objective of this paper is to apply ML algorithms to classify breast cancer outcomes using a small publicly available data set.

1.2 Motivation of the Project:

As of 2021, on average, 1 in 8 women (approx. 12%) would develop invasive breast cancer at some point during her life.

5-year survival rate for breast cancer is 100% with early detection and 15% with late detection (Cancer research)

Machine learning (ML) techniques play a key role in healthcare in recent years.

In the case of breast cancer, machine learning techniques can be used to distinguish between malignant and benign tumours for enabling early detection.

Most ML based applications focus on large data sets citing ML’s ability to handle big data.

However, from a user’s perspective most users have access to publicly available small data sets.

Thus, it is interesting to analyse if the traditional non complex basic ML algorithms can achieve high accuracy classifications using small datasets.

1.3 Basic Description of the Project:

This project is based on “new evolutionary learning approaches for feature selection on biomedical data”. In this project we have used machine learning skills using python programming language .We have used different libraries of python which helps us to apply machine learning algorithms in a much effective manner.

Machine learning is :-

•a branch of artificial intelligence

•employs a variety of statistical, probabilistic and optimization techniques

•allows computers to “learn” from past examples

•detect hard-to-discern pattern from large, noisy or complex data .

The goal of ML is to develop algorithms which can learn and improve over time and can be used for predictions. Most ML researchers concentrate on automatic machine learning (aML), where great advances have been made.

Particularly, probabilistic ML is extremely useful for health informatics, where most problems involve dealing with uncertainty.

Here , we build automated models those have high accuracies in detecting and classifying benign cells and malignant cells thus greatly easing the process of diagnosing breast cancer nowadays.

2. LITERATURE REVIEW

Review of Related Work

Segmentation :

Segmentation is the process of dividing the image into meaningful regions. The two types of segmentation are supervised and unsupervised segmentation. In supervised learning, the user should have knowledge about the number of classes and in unsupervised learning, the user may not have knowledge about the number of classes.

Method I - Pegah Faridi proposed an algorithm for the automated detection of the abnormal masses present in the breast tissues. Bilateral filtering is used as the preprocessing step and edges are preserved while applying the filter. The algorithm segments the image to the centroid of the nucleus and boundary of the cell nucleus. Steps like Gamma correction and thresholding, a morphological operation like dilation and erosion, DoG filtering, and thresholding are applied on the bilateral filtered image to detect the centroid of the nucleus. The cell nucleus boundary is identified by finding the points of change in the brightness level. The average accuracy in finding the cell nucleus and cell boundaries are 86% and 87% respectively. The distinguishing feature of this algorithm is no need for training which is considered to be the most time-consuming process.

Method II - G.Bharatha Sreeja initiated an algorithm for the detection of tumors in the breast tissue. The author has used linear contrast stretching as the preprocessing step. In this paper, the author has done coarse and fine segmentation of the mammogram. For doing the coarse segmentation apply one-dimensional wavelet form of the PDF of wavelet applied images of various channels and compute the local minima. The local minimum is fixed as the threshold value. For performing the fine segmentation form small and large windows. Compute the mean, maximum, and minima of the window for calculating the threshold value. Also, compute the difference between the minimum and the maximum. If the gray level value of the pixel is greater than the threshold and if the difference is greater than the mean value then the particular pixel belongs to the abnormal region, otherwise the pixel belongs to the normal region.

Method III - Albert Gubern developed an algorithm for the automated segmentation of the breast and estimating its density using MRI. As an initial step for segmenting the breast it is separated from the other organs of the body like the heart, lungs, thorax, etc. Bias field correction and image normalization are carried out as the preprocessing step. From the MRI image sternum forms the main identification for separating the breast from the body. From the segmented breast, the dense parts are segmented. The EM algorithm is used to segment the fibroglandular tissue. Different threshold methods were used for the segmentation of the fibroglandular tissue. The use of a single threshold showed heavy asymmetry between the left and right breast. This asymmetry is reduced by N3 bias field correction and segmenting the two breasts separately.

Method IV - Hadjidj Ismahan initiated an algorithm for segmenting the mass present in the breast using mammogram. The mammogram image is passed through a morphological filter to remove unwanted radiopaque artifacts present in the image. The Image enhancement process is carried out to enhance the features present in the image. Then the image is converted to a binary image. The breast alone is obtained by convolving the mask obtained by the morphological process and the filtered image. The masses present in the image are identified by the watershed algorithm. For evaluating the performance of the suggested algorithm area overlap parameter is found.

Method V - Amresh Nikam proposed an algorithm for the segmenting of the nucleus of the breast cells from histology images. The input images undergo histogram equalization for making the darker images darker and lighter images lighter. The image is converted to grayscale images. Average of Potent and applied on the gray image to remove the pixels which are more close to the background. Difference method, Minimum error method, Otsu method, Max entropy method are computed. Thus the pixels which are whiter are identified and they are superimposed on the histology images. After this, graph-based segmentation is applied to the superimposed image. The next step is eliminating the large region by overlapping the original histology image and white pixel obtained from the segmentation. Hard thresholding, is performed on the image to neglect the false region and exactly identify the nuclei. In the next step morphological operations like dilation and erosion are performed on the image.

Feature Extraction and Classification:

Rafael Lobet proposed five different feature extraction techniques and compared the

performance of these methods. The feature extraction methods proposed are :

Gray Map: The gray map values of a pixel along with its neighbourhood followed by Principal Component Analysis are computed for every pixel in the image.

Sobel: This filter consists of two kernels that can be used for detecting the magnitude and direction of the image. However, the author has used Sobel filters to detect the magnitude only.

Spatial Gray Level Dependence Matrices (SGLDM): The features considered are Difference Entropy, Sum Variance, Entropy, Difference Variance, Angular Second Moment, Contrast, Correlation, Variance, Inverse Difference Moment, Sum Average, and sum entropy. For the calculation of SGLDM, four angle and 1 to 4-pixel distance is considered.

Average Fraction Under the Minimum (AFUM): The pixels with gray level values less than the gray level value of the pixel pij within a radius r is calculated. The average of all these pixels is computed. The outstanding feature of this method is avoiding the time-consuming training process.

5) SFUM: This method is similar to the AFUM. The only difference is that the average of all the minimum pixels is not calculated. Instead, all the minimum value pixels are maintained as features. PCA is computed to reduce the dimension of this huge database. For these five feature extraction methods, K nearest neighbour is used as the classifier. Gray map and SFUM showed promising results compared to the other methods. P.Shanthakumar proposed to review various MR Images to classify and segment

the different image sets.

3. PROPOSED MODEL/ ALGORITHMS

3.1 Logistic Regression :

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

****

where x = independent variable

y = dependent variable(function of x)

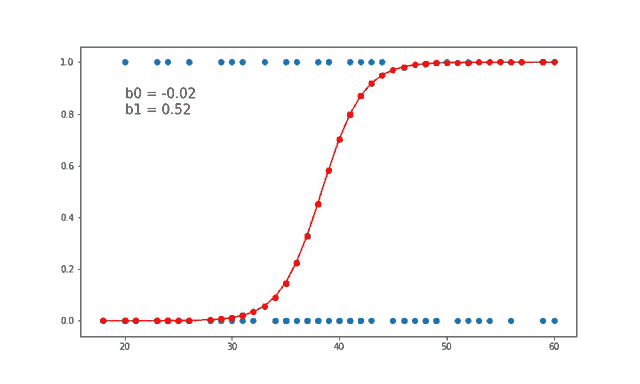


Fig. 3.1 Logistic Regression Graph

3.2 Decision Tree :

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

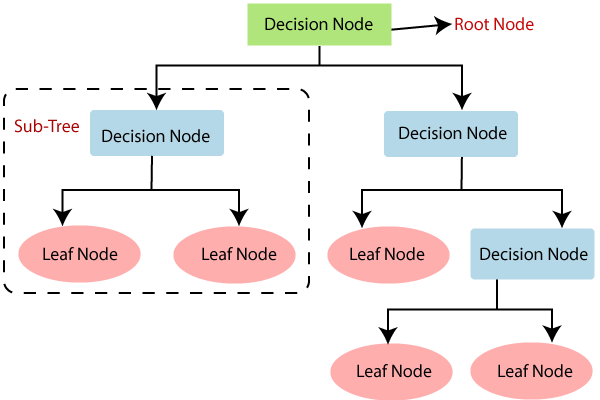
****

Fig. 3.2 Decision Tree Process

3.3 Feature Selection (Filtering) :

While building a machine learning model for real-life dataset, we come across a lot of features in the dataset and not all these features are important every time. Adding unnecessary features while training the model leads us to reduce the overall accuracy of the model, increase the complexity of the model and decrease the generalization capability of the model and makes the model biased. Even the saying “Sometimes less is better” goes as well for the machine learning model. Hence, feature selection is one of the important steps while building a machine learning model. Its goal is to find the best possible set of features for building a machine learning model.

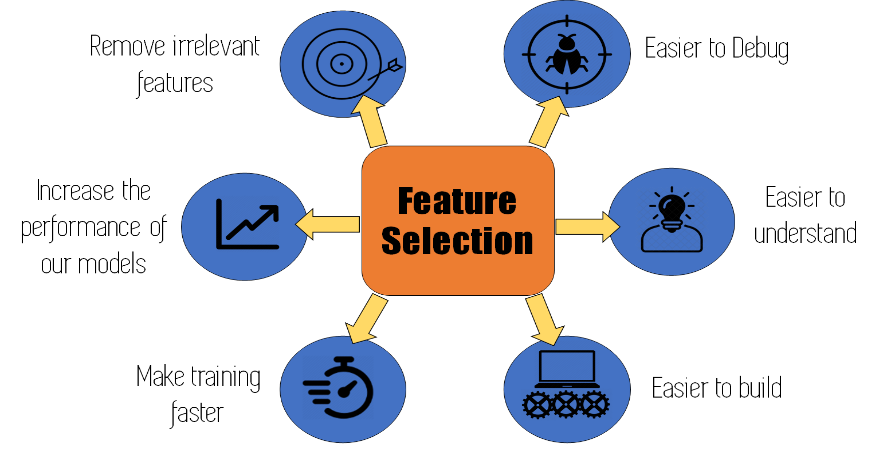
****

Fig. 3.3 Feature Selection Uses

Filter Methods : These methods are generally used while doing the pre-processing step. These methods select features from the dataset irrespective of the use of any machine learning algorithm. In terms of computation, they are very fast and inexpensive and are very good for removing duplicated, correlated, redundant features but these methods do not remove multicollinearity. Selection of feature is evaluated individually which can sometimes help when features are in isolation (don’t have a dependency on other features) but will lag when a combination of features can lead to increase in the overall performance of the model**.**

****

3.4 Random Forest :

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept ofensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

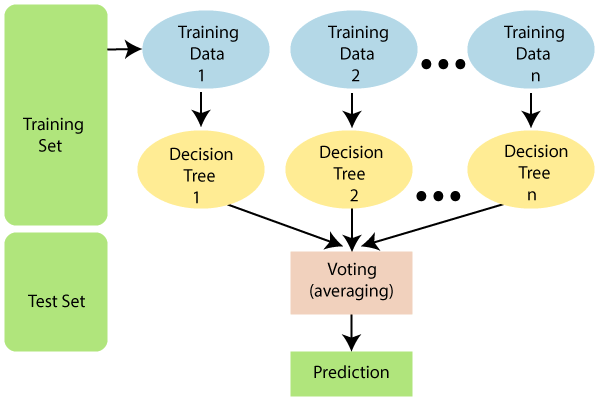
****

Fig. 3.4 Random Forest Classifier

4. SIMULATION AND RESULTS

4.1 Experimental Set Up :

Platform used :– Jupyter Notebook

Programmedin **:–** Python

Librariesused **:–**

Numpy (used for performing mathematical operations on our data).

Pandas (used for data manipulation and analysis).

Matplotlib (used for graphical representation of data).

Scikit Learn (used for model training and accuracy measurement).

Seaborn (used for Data visualization).

Workflow Process :

Loading Libraries and Dataset – Python libraries are imported and using pandas , the dataset is loaded as a dataframe.

Feature selection from the dataset – Filtering is done to remove the null values from all the rows and columns. And we got a new fresh dataset.

Value Counts – From the ‘diagnosis’ feature , total no of benign and malignant cases are counted. We see that our dataset is imbalanced having more no of benign than malignant.

Data Visualization – Using Seaborn libraries, the categorical values are visualized in the form of graphs.

Data preprocessing – Using scikit learn, the fresh data is preprocessed and labelled into two binary values (1 for cancerous or malignant and 0 for benign). Then correlation heatmap is plotted for each pair of features to visualize the impact of the features.

Splitting Dataset – Using scikit learn ,75% data is splitted for training the model and rest 25% for testing purpose.

Model Training – The model is trained using 3 different supervised learning algorithms(Logistic Regression, Decision Tree, Random Forest). Accuracy is calculated for the training models. We found Decision Tree gave the highest accuracy in model training.

Confusion Matrix – From sklearn , confusion matrix is formed for 3 algorithms and testing accuracy is calculated . Here we found Random Forest gave the highest accuracy over the other two.

Random Forest Prediction – Finally we preferred Random Forest Classifier for Breast Cancer Prediction and once again predict the result for our test datasets and and get it in binary formats.

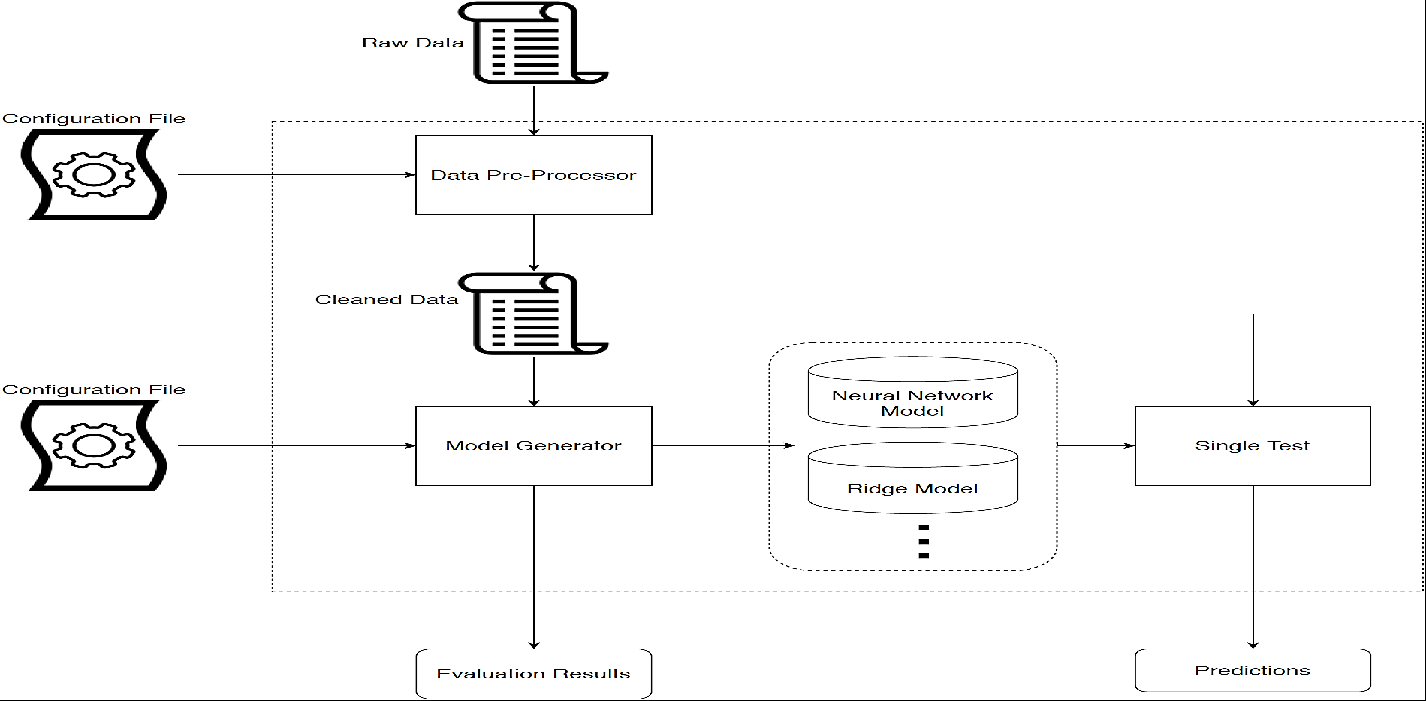


Fig. 4.1 Workflow Process

4.2 Experimental Results:

Code Analysis and Respective Outputs

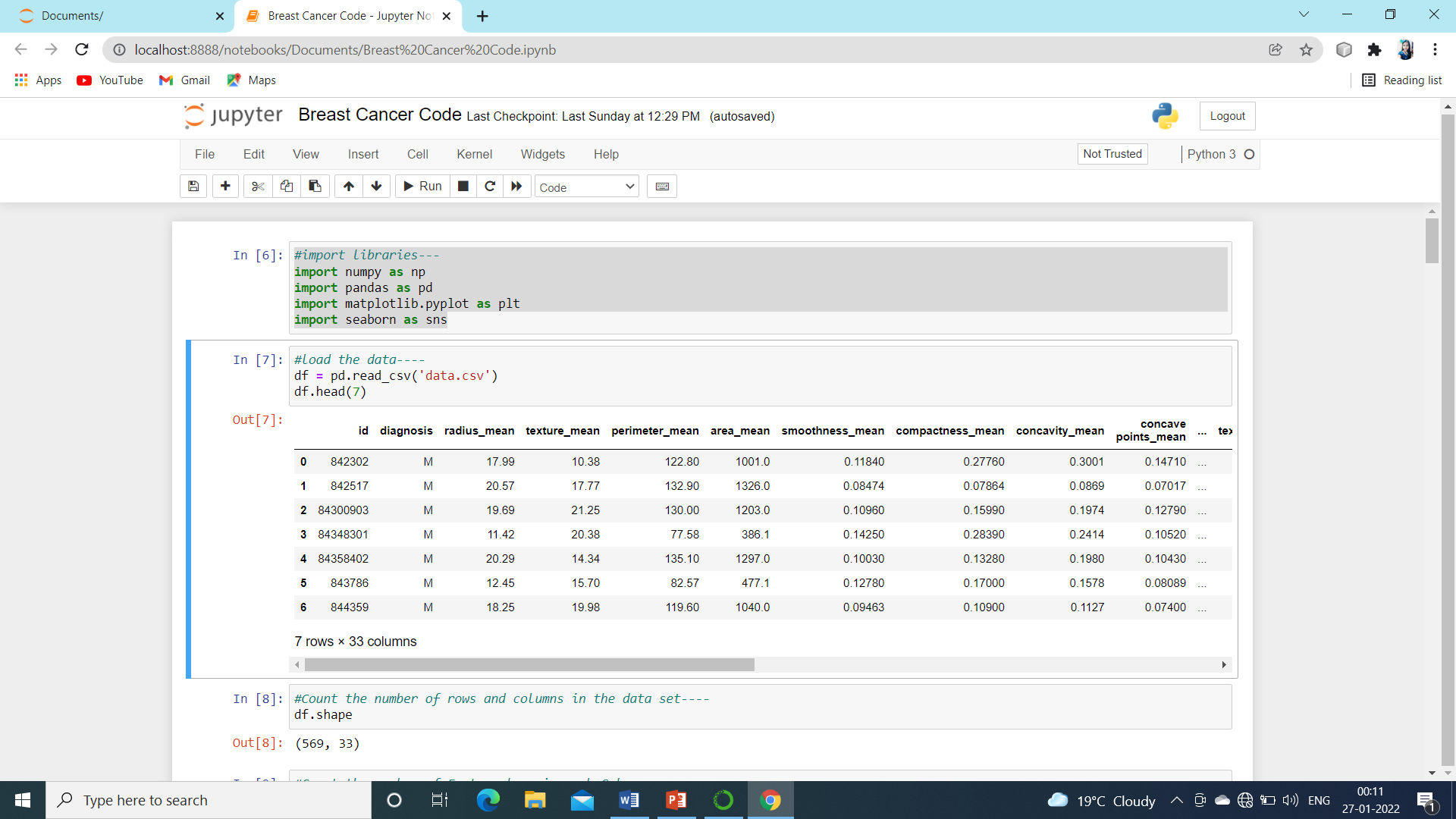


Fig. 4.2.1 Module import, Dataset Loading

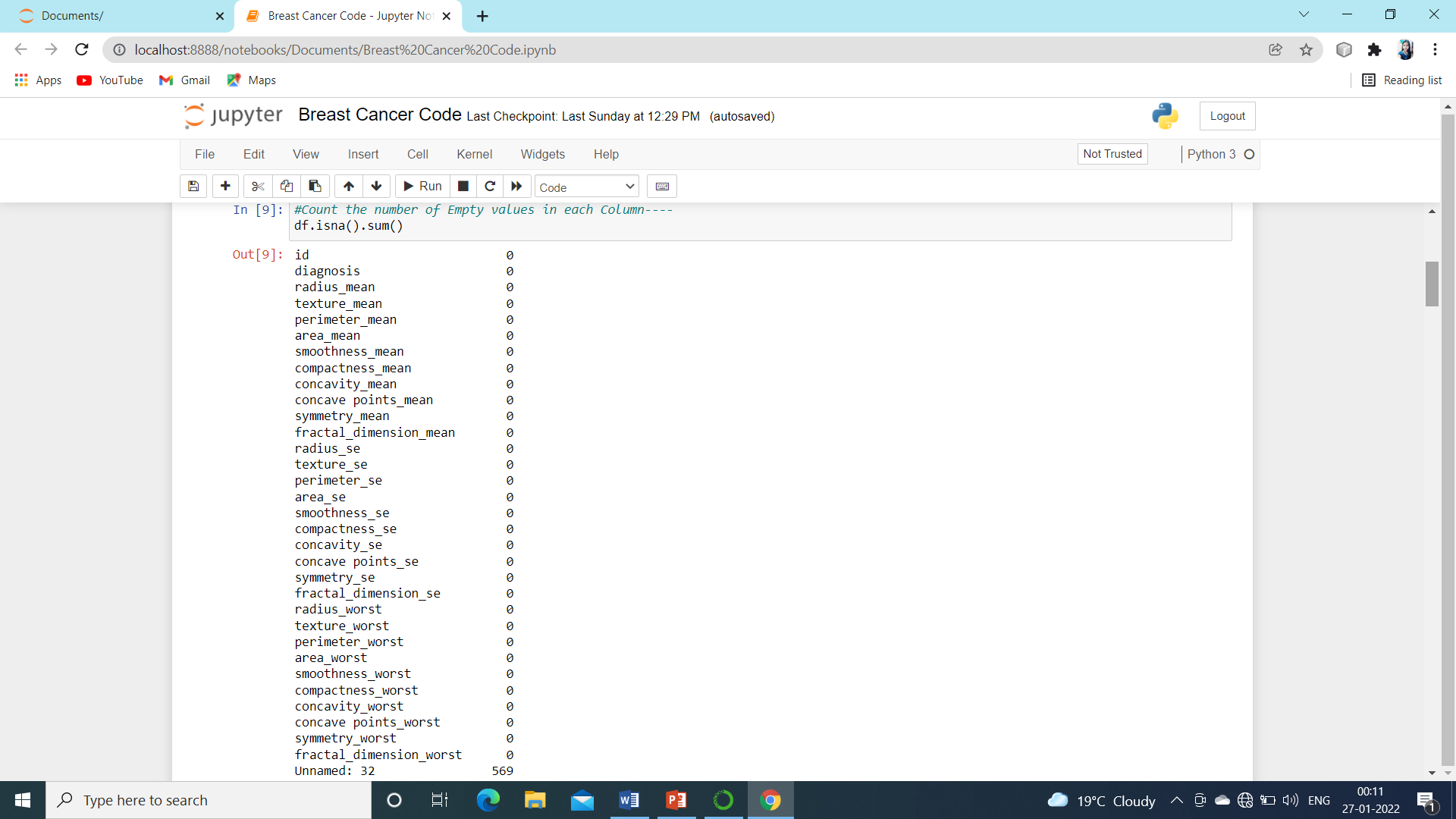


Fig. 4.2.2 Feature Selection, Checking for Null values

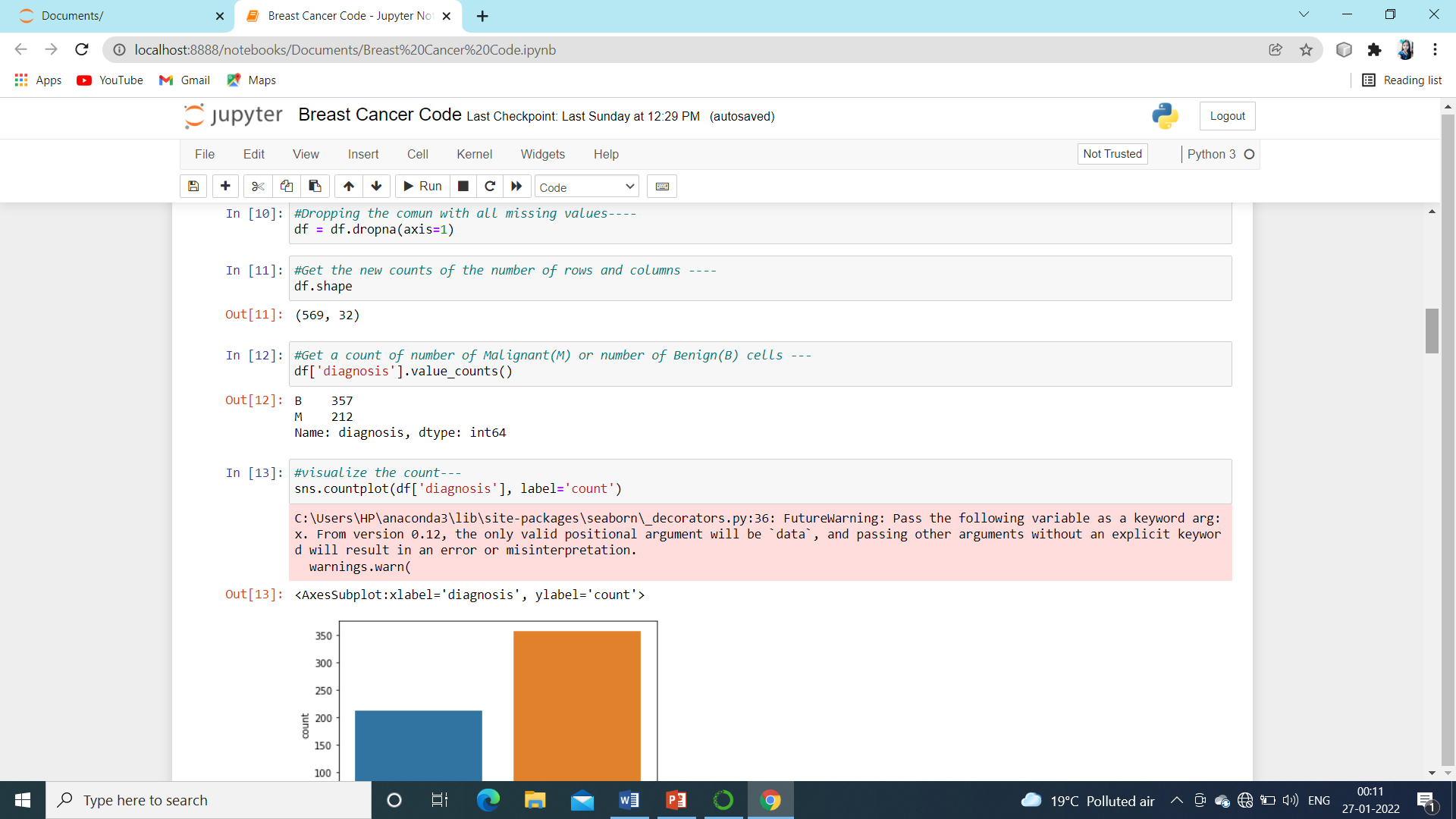


Fig. 4.2.3 Data Cleaning and Counting classifying variables

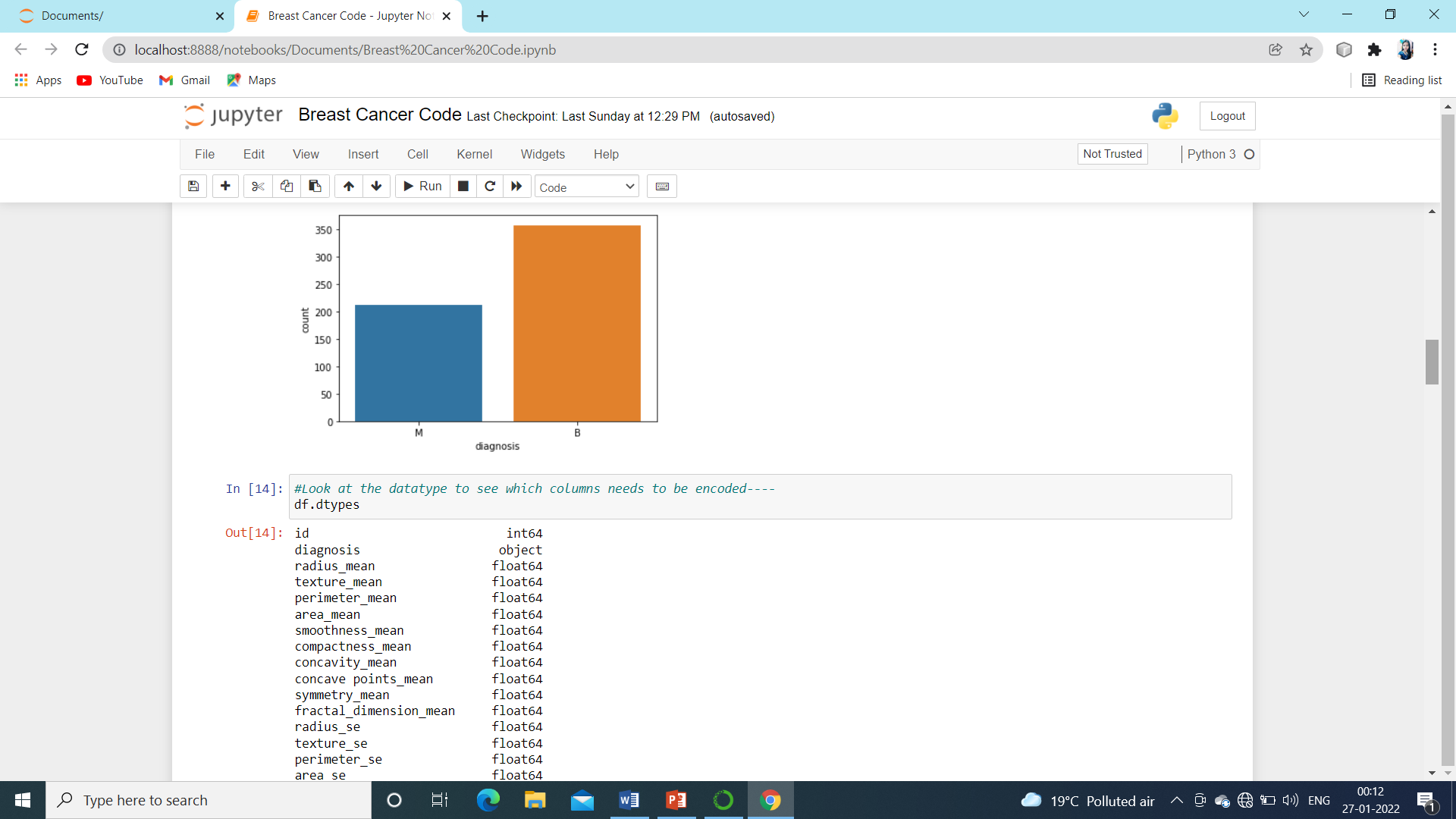


Fig. 4.2.4 Plotting Bars Graphs for Classifying Variables

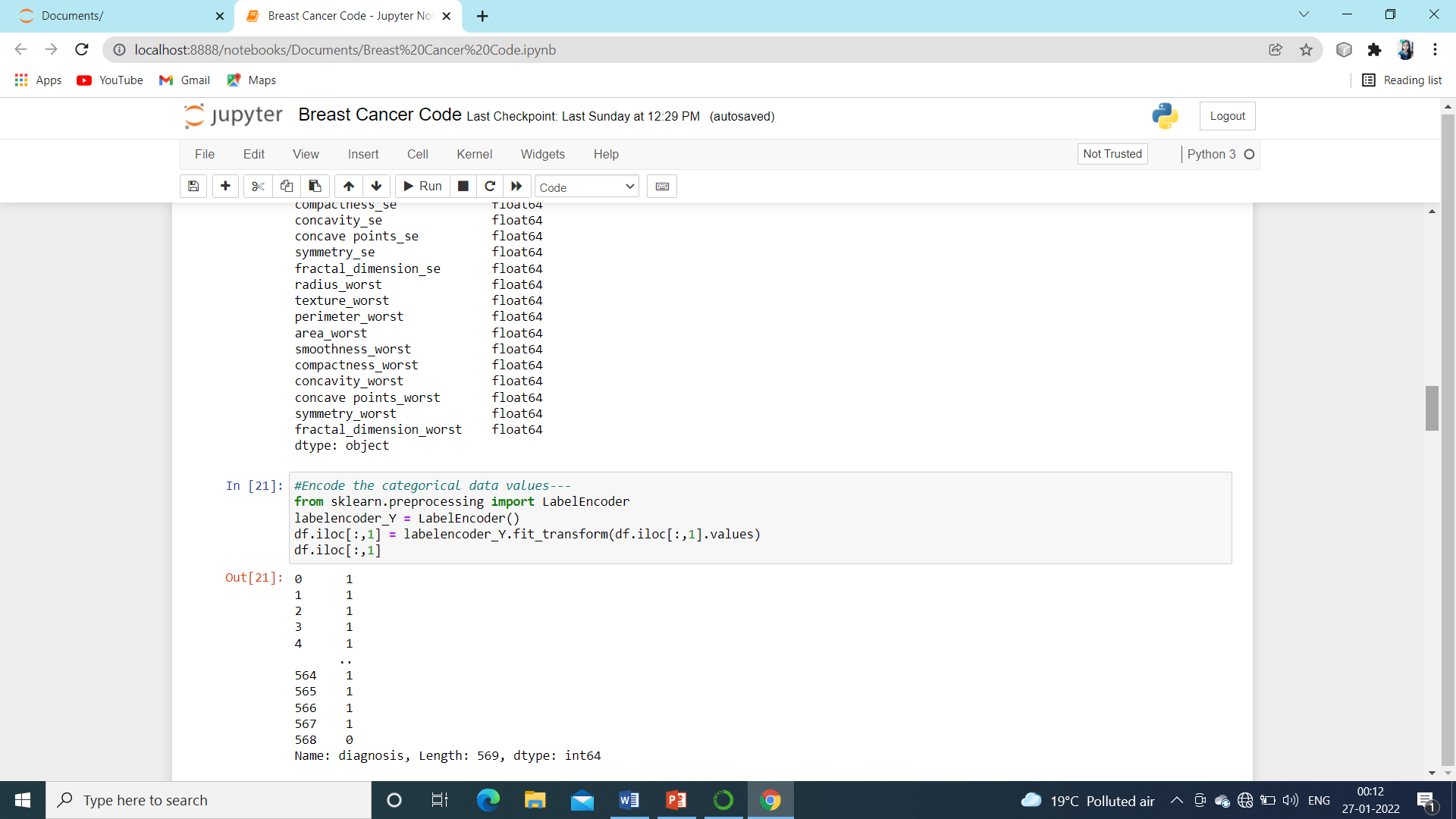


Fig. 4.2.5 Data Preprocessing

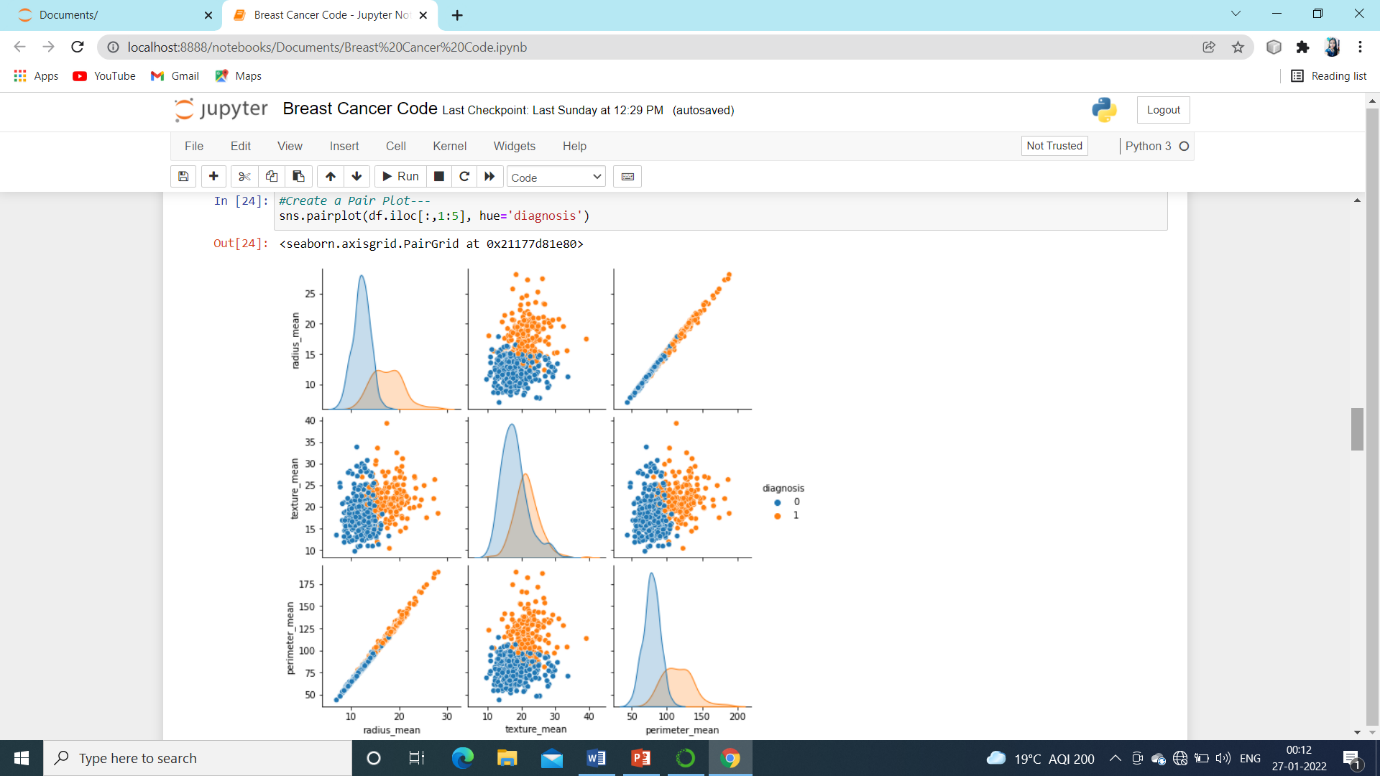


Fig. 4.2.6 PlairPlot is plotted against Each pair of Features

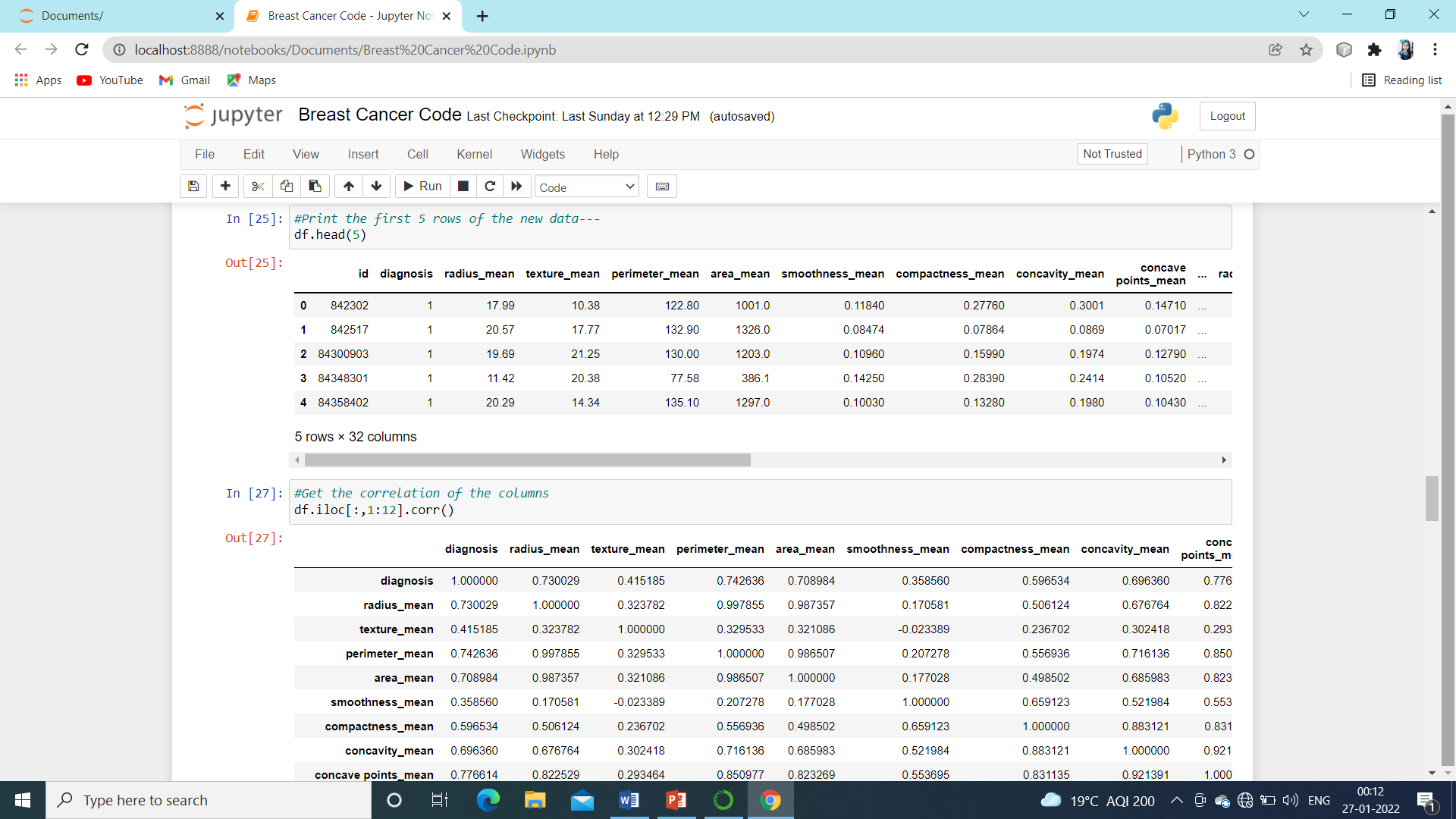


Fig. 4.2.7 Correlation Matrix

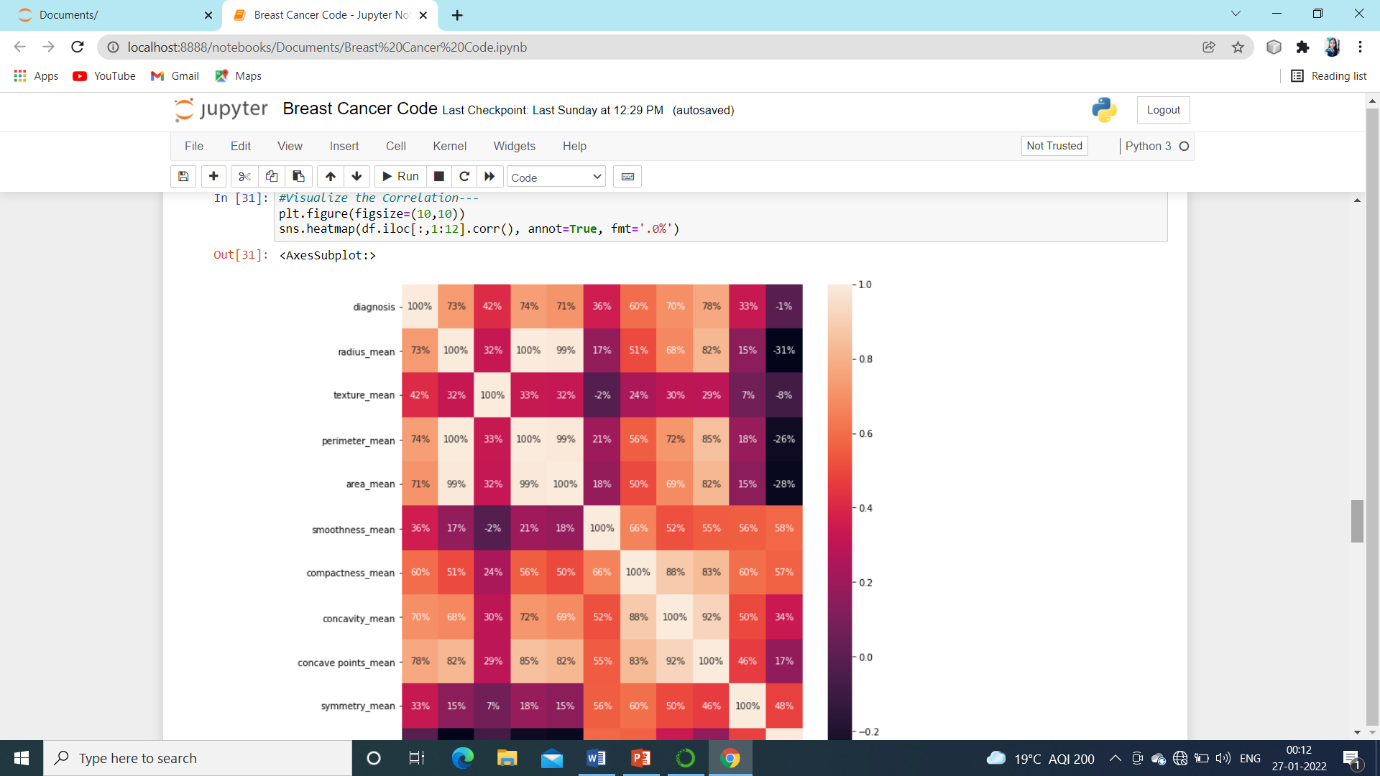


Fig. 4.2.8 Heatmap on Correlation Matrix

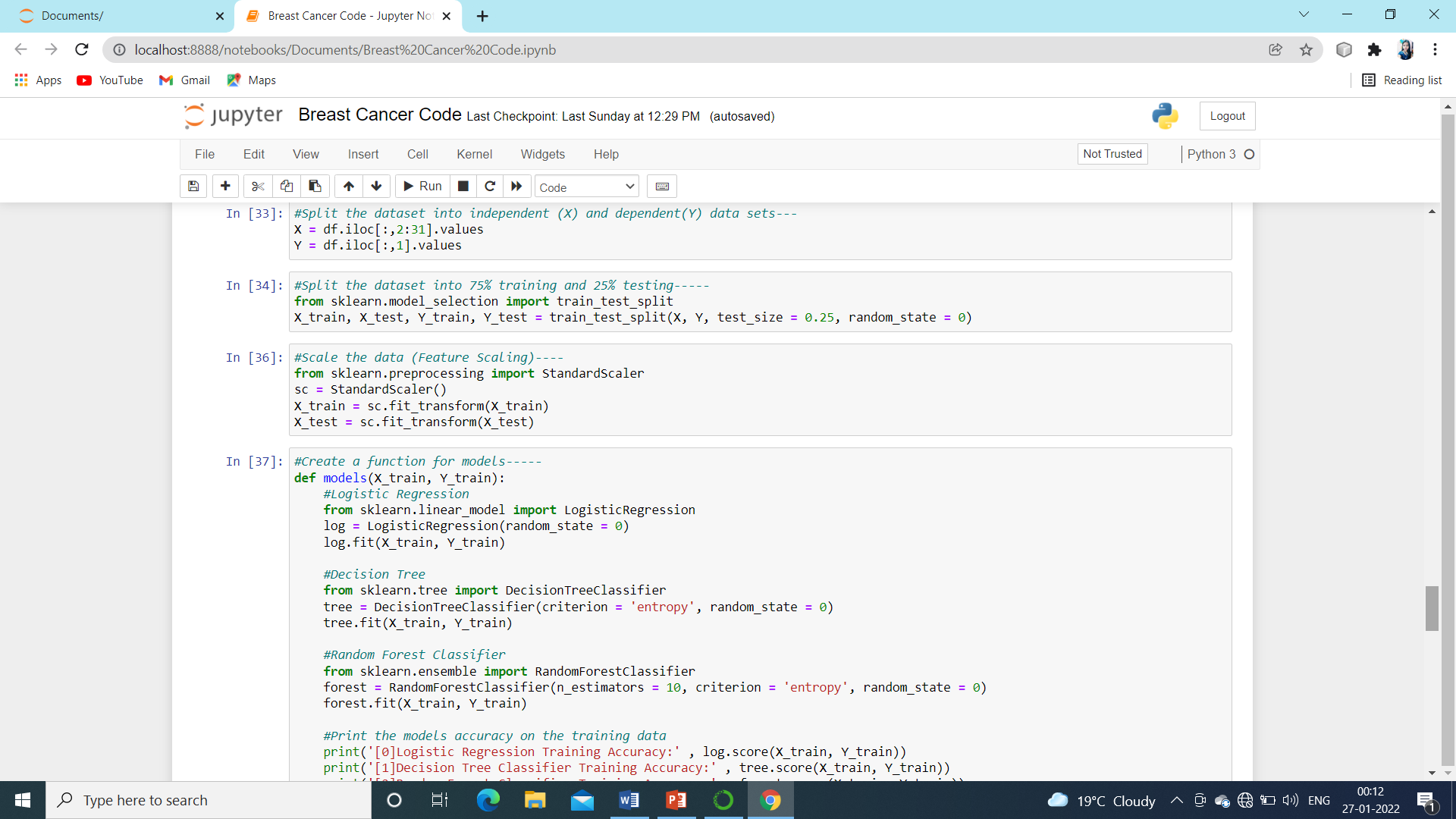


Fig. 4.2.9 Splitting Dataset into Training and Test dataset

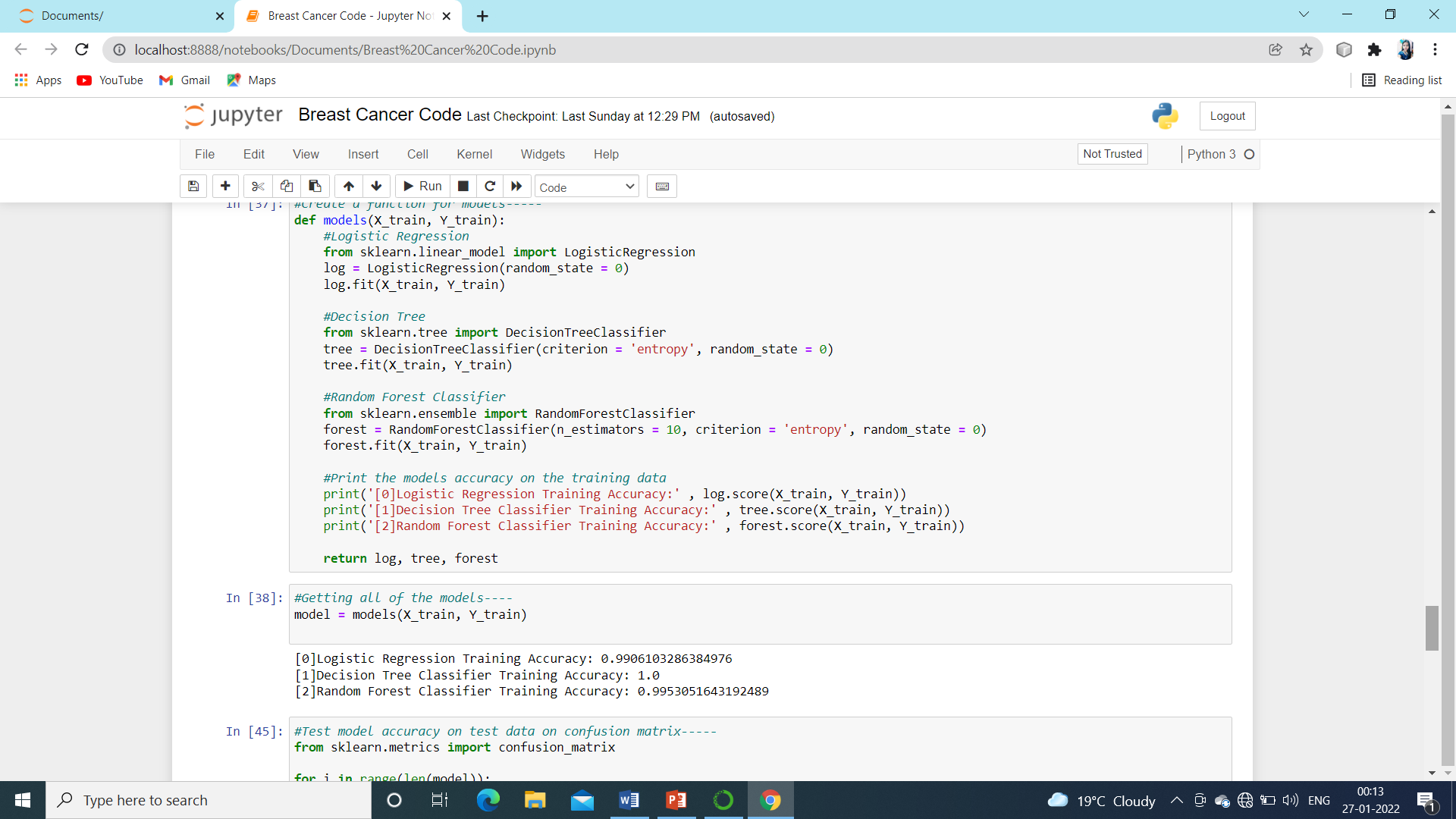


Fig. 4.2.10 Model Training on Training Dataset

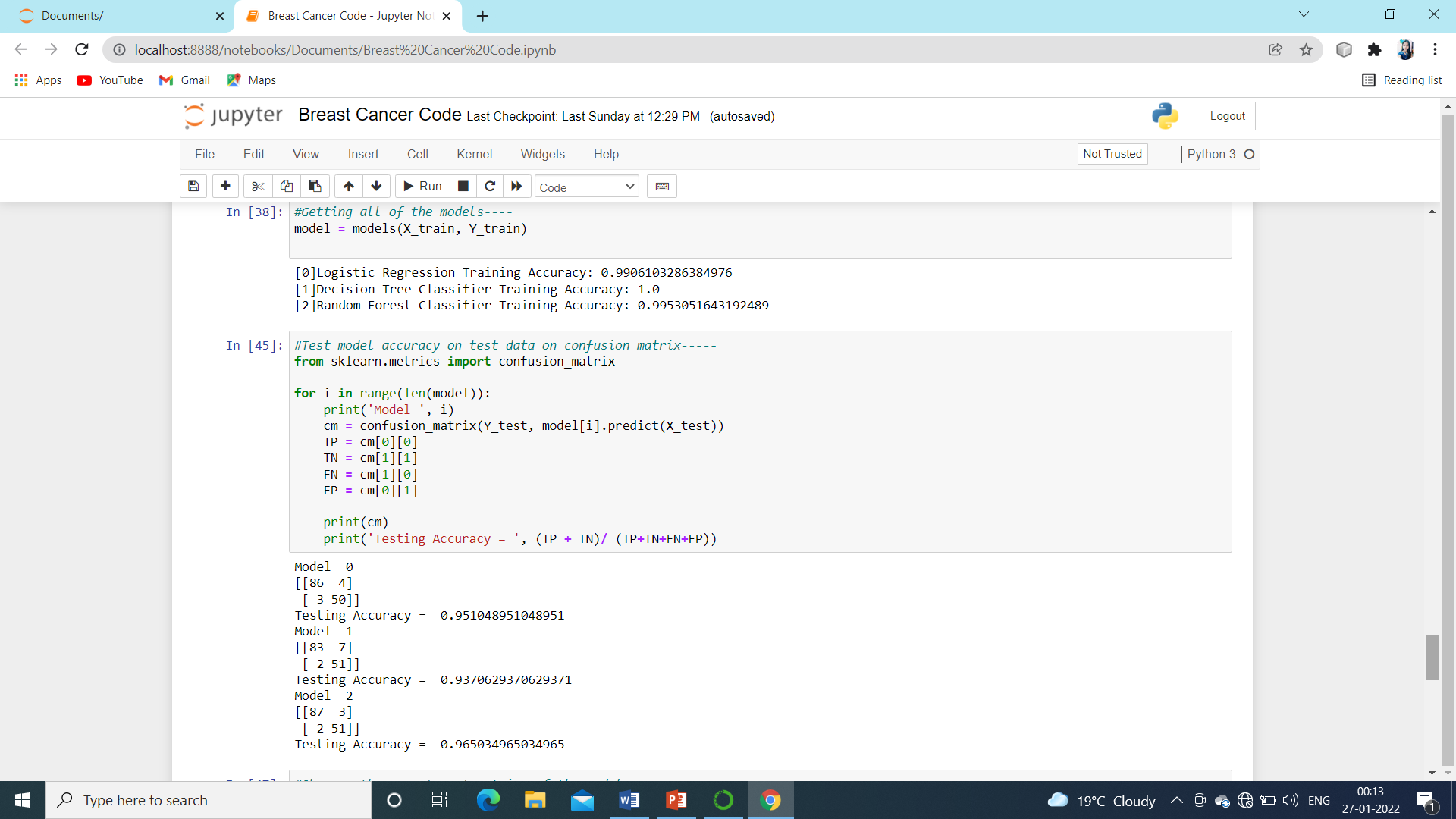


Fig. 4.2.11 Confusion Matrix and Accuracy

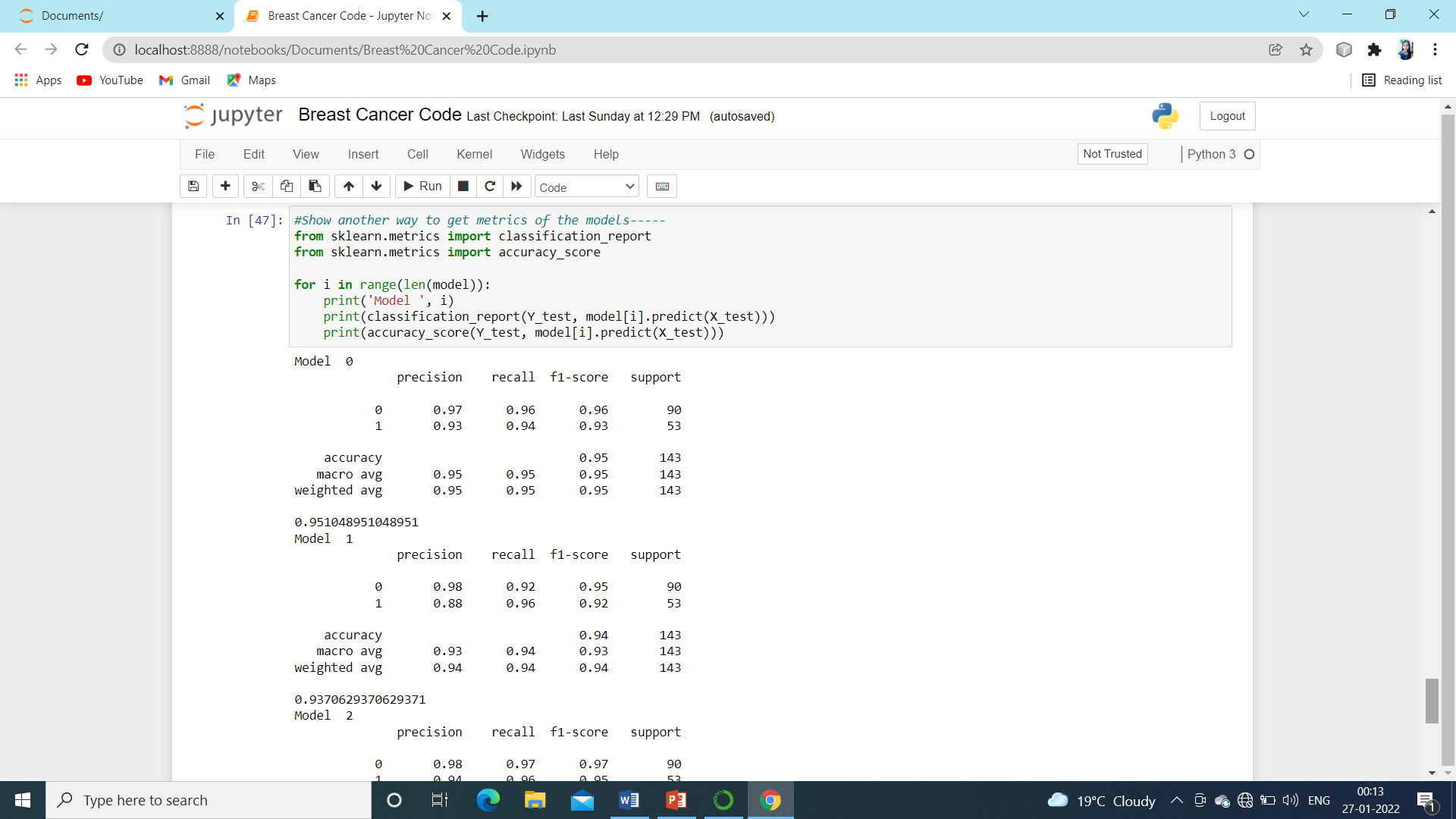


Fig. 4.2.12 Precision

Accuracy and Results:

From the above implementation , we found that Random Forest Classifier is having the highest accuracy of 96.50% over the other two algorithms used here for testing data.

Thus making the prediction precise to about 98% of the expected values.

5. DISCUSSION AND CONCLUSION

5.1 Discussion :

Three foci of Breast Cancer Prediction and Prognosis:

The prediction of breast cancer susceptibility – risk assessment prior to occurrence (Diagnosis).

The prediction of breast cancer recurrence– likelihood of redeveloping (Prognosis).

The prediction of breast cancer survivability – life expectancy, survival, progression, tumor-drug sensitivity (Prognosis) .

The success of Prognosis prediction is dependent on the quality of the Diagnosis.

Pros and Cons of Machine Learning Algorithms:

Logistic Regression – Causal effect

DecisionTree – English rule, segmentation, variable selection, use both categorical and interval with missing values.

RandomForest – reduce overfitting.

5.2 Future Work:

MachineLearningis a key technique to automate Medical disease classification with much improved architecture. Further tests and research are needed.

Further specification: SVM (linear, polynomial, RBF, sigmoid kernel) .

Methods: Clustering, segmentation, two stage modelling, cross validation .

Data: subsets (HER2+/-), different cancers, unstructured data .

Architecture: HDFS Laser server, In-Memory statistics, Results Visualization.

5.3 Conclusion:

This study identifies Random Forest model to be most successful in breast cancer classification. (96.5% accuracy rate, and the least number of fault predictions)

concave\_points\_worst and perimeter\_worst are the most important features for classifying breast cancer outcomes.

Our results show that ML algorithms can classify breast cancer outcomes with high accuracy and identify key characteristics even for small datasets.

We highlight the significant potential in using ML techniques as a diagnostic tool for early detection of breast cancer in general.

REFERENCES

Coursera

Youtube

Geeksforgeeks

<https://scikit-learn.org/stable/>